

EUROPEAN PATENT OFFICE

Patent Abstracts of Japan

PUBLICATION NUMBER : 08320768
PUBLICATION DATE : 03-12-96

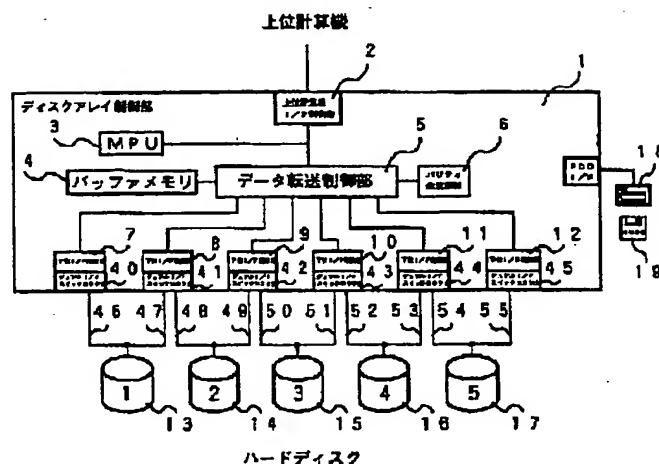
APPLICATION DATE : 26-05-95
APPLICATION NUMBER : 07127761

APPLICANT : HITACHI LTD;

INVENTOR : IWASAKI HIDEHIKO;

INT.CL. : G06F 3/06 G06F 3/06

TITLE : DISK ARRAY DEVICE



ABSTRACT : PURPOSE: To provide the disk array device which is placed in continuous operation if a fault occurs to one slave I/F control part or the cable connecting the slave I/F control part and a hard disk.

CONSTITUTION: Between slave I/F control parts 7-12 of a disk array control part 1 and hard disks 13-17, dual I/F connectors 40-45 are arranged which switch the main path using cables 47, 49, 51, 53, and 55 used when the disk array device is in normal operation as communication paths to a standby path using cables 46, 48, 50, 52, and 54 used if a fault occurs to one slave I/F control part or the cable connecting the slave I/F control part and a hard disk as communication paths; and each hard disk has two communication means for the slave I/F control parts 7-12.

COPYRIGHT: (C)1996,JPO

(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号

特開平8-320768

(43) 公開日 平成8年(1996)12月3日

(51) Int.Cl. ⁶	識別記号	庁内整理番号	F I	技術表示箇所
G 0 6 F 3/06	5 4 0		G 0 6 F 3/06	5 4 0
	3 0 4			3 0 4 H

審査請求 未請求 請求項の数10 O L (全 24 頁)

(21) 出願番号 特願平7-127761

(22) 出願日 平成7年(1995)5月26日

(71) 出願人 000005108

株式会社日立製作所

東京都千代田区神田駿河台四丁目6番地

(72) 発明者 松本 純

神奈川県川崎市麻生区王禅寺1099番地株式

会社日立製作所システム開発研究所内

(72) 発明者 大枝 高

神奈川県川崎市麻生区王禅寺1099番地株式

会社日立製作所システム開発研究所内

(72) 発明者 市川 正敏

神奈川県川崎市麻生区王禅寺1099番地株式

会社日立製作所システム開発研究所内

(74) 代理人 弁理士 小川 勝男

最終頁に続く

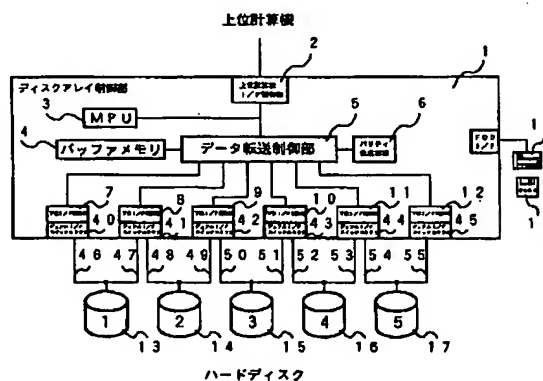
(54) 【発明の名称】 ディスクアレイ装置

(57) 【要約】

【目的】 1台の下位I/F制御部または下位I/F制御部とハードディスクを接続するケーブルの故障時に、連続運用可能なディスクアレイ装置を提供する。

【構成】 ディスクアレイ制御部1の下位I/F制御部7～12とハードディスク13～17の間に、ディスクアレイ装置の正常運用時に用いケーブル47、49、51、53、55を通信経路とする主経路から、1台の下位I/F制御部または下位I/F制御部とハードディスクを接続するケーブルの故障時に用いケーブル46、48、50、52、54を通信経路とする予備経路に切替えるデュアルI/Fコネクタ40～45を備え、各々のハードディスクが下位I/F制御部7～12に対して2本の通信手段を持つ。

図 1



(2)

特開平8-320768

1

【特許請求の範囲】

【請求項1】上位に上位計算機、下位に複数のハードディスクを接続し、内部に上位計算機との間でコマンド及びデータの授受を行う上位計算機I/F制御部、複数のハードディスクとの間でコマンド及びデータの授受を行う下位I/F制御部、コマンドのアドレス変換、データの集合分散制御を行うデータ転送制御部、データからパリティの生成を行うパリティ生成回路、上位計算機と複数のハードディスクの間で転送するデータを一時的に格納保持するバッファメモリ、さらに以上の各構成要素を統括制御するMPUを内部に持つディスクアレイ制御部において、N台の下位I/F制御部+1台の予備用の下位I/F制御部を持つことで、1台の下位I/F制御部の故障発生時に連続運用が可能なことを特徴とするディスクアレイ装置。

【請求項2】請求項1において、1台の下位I/F制御部の故障発生時に、下位I/F制御部とハードディスク間の通信経路を正常運用時に使用する主経路から、下位I/F制御部及びディスクアレイ制御部とハードディスクの間を接続するケーブルの故障発生時に使用する予備経路に切替えることで、連続運用が可能なディスクアレイ装置。

【請求項3】請求項2において、各下位I/F制御部と各ハードディスク間に正常運用時に使用する主経路と、下位I/F制御部及び下位I/F制御部とハードディスクを接続するケーブルの故障発生時に使用する予備経路を切替る通信経路切替スイッチを付加したディスクアレイ装置。

【請求項4】請求項3において、1台の下位I/F制御部の故障発生時にディスクアレイ制御部を構成する基盤全体ではなく、下位I/F制御部単位で故障した下位I/F制御部の交換を可能とするため、各々の下位I/F制御部をソケットにより抜き差し可能な構造を持つディスクアレイ装置。

【請求項5】請求項4において、外部機器専用の下位I/F制御部を持たずにMT、MO等のバックアップ機器を接続可能なディスクアレイ装置。

【請求項6】請求項5において、1台の下位I/F制御部の故障発生時にも残りの下位I/F制御部を利用してMT、MO等のバックアップ機器にデータバックアップ可能なディスクアレイ装置。

【請求項7】請求項4に記載の通信経路切替スイッチを搭載し、従来のディスクアレイ装置の下位I/F制御部とハードディスク間のケーブルの途中に接続されるデータボード。

【請求項8】請求項8に記載のデータボードを、従来のディスクアレイ装置の下位I/F制御部とハードディスク間に付加したディスクアレイ装置。

【請求項9】上位に上位計算機、下位に複数のハードディスクを接続し、内部に上位計算機との間でコマンド及

2

びデータの授受を行う上位計算機I/F制御部、複数のハードディスクとの間でコマンド及びデータの授受を行うシリアルI/F制御部、コマンドのアドレス変換、データの集合分散制御を行うデータ転送制御部、データからパリティの生成を行うパリティ生成回路、上位計算機と複数のハードディスクの間で転送するデータを一時的に格納保持するバッファメモリ、さらに以上の各構成要素を統括制御するMPUを内部に持つディスクアレイ制御部において、ディスクアレイ制御部のシリアルI/F制御部をデュアル化したことを特徴とするディスクアレイ装置。

【請求項10】請求項9において、2台のシリアルI/F制御部と各ハードディスク間に正常運用時に使用する主経路と、下位I/F制御部及び下位I/F制御部とハードディスクを接続するケーブルの故障発生時に使用する予備経路を切替る通信経路切替スイッチを付加したディスクアレイ装置。

【発明の詳細な説明】

【0001】

【産業上の利用分野】本発明は、ハードディスク、光ディスク、半導体ディスクなどのディスク装置を複数使用するディスク装置システムにおけるディスクアレイ装置に関する。

【0002】

【従来の技術】ファイルシステムの大容量化と高いランザクション性能が要求される中、高いデータ処理性能と信頼性を併せ持つ方式として、米国特許US4870643号に開示されるパラレルドライブアレイ・ストレージシステムがある。本方式は従来のディスクアレイ装置の基本構成を示すものである。ここで従来のディスクアレイ装置について図16を用いて説明する。

【0003】ディスクアレイ装置は主に上位計算機から受けたリード/ライトコマンドにより、上位計算機とハードディスクの間でデータの分散集合制御を行うディスクアレイ制御部1と上位計算機が転送したデータを格納するハードディスク13～17で構成される。またデータ保持の信頼性を高める意味で、ハードディスク13～17のデータをバックアップするMTやMO等のバックアップ機器をオプションとして接続することが多い。ディスクアレイ制御部1は上位計算機との間でコマンド及びデータの授受を行う上位計算機I/F制御部2、各ハードディスク13～17及びバックアップ機器21との間でコマンド及びデータの授受を行う下位I/F制御部7～12、上位計算機I/F制御部2と下位I/F制御部7～11の間でデータを転送制御するデータ転送制御部5、ハードディスク13～17に新しいデータをライティングの際にデータに付加するパリティを生成するパリティ生成回路6、また上位計算機とハードディスク13～17間のデータ転送の際に、一時的にデータを保持するバッファメモリ4、さらにディスクアレイ制御部1全体

(3)

特開平8-320768

3

を統括制御するMPU3等で構成される。本ディスクアレイ装置は各ハードディスク13~17それぞれに下位I/F制御部7~11を備えるため、複数ハードディスク13~17のバラレルなリード/ライト動作が可能となり、単体のハードディスクに比べて特にアクセス性能に優れているという特徴を持つ。また、データをハードディスクにライトする際に、パリティ生成回路6が図17に示すようにデータの排他的論理和からパリティを生成し、データと共にハードディスクのある場所にライトするため、例えば、3番目のハードディスク15に故障が発生しても、残りのハードディスク13、14、16、17のデータD1、D2、D4、D5及びパリティPから喪失データD3を再生することができる。

【0004】ハードウェア構成を持つディスクアレイ装置は、下位I/F制御部7~12に特にSCSI<Small Computer System Interface>を用いたもので、SCSIはコンピュータ周辺機器の一般的なI/F制御部として広く普及している。しかし、近年MPU、メモリ及びハードディスク等のハードウェアの性能の上昇や小型化が著しいため、SCSIの持つデータ転送速度や転送距離による制約、大型のI/Fコネクタが問題となっており、これに変わるものとしてシリアルI/F制御部が注目されている。図18はハードディスク24~28とディスクアレイ制御部29の間にシリアルI/F制御部22を用いたディスクアレイ装置を示した図である。ディスクアレイ制御部29を構成する要素はシリアルI/F制御部22を除いてSCSI I/F制御部を用いたディスクアレイ装置と同様であるが、シリアルI/F制御部22の高い転送能力を引き出すために、専用のシリアルI/F制御用MPU23を持っている。またハードディスク24~28は各々のデータ入出力口に、ディスクアレイ制御部29同様にシリアルI/F制御部30~34を備えている。図19はディスクアレイ制御部29側のシリアルI/F制御部22（特にSSA<Serial Storage Architecture>）とハードディスク24~28側のシリアルI/F制御部30~34の詳細構成を示している。シリアルI/F制御部22、30~34はデータの送信及び受信ポートを2ポートずつ備えており、各シリアルI/F制御部22、30~34がループ状に連なることで、隣り合うシリアルI/F制御部が双方向の通信経路を持つことができる。このことから例えば図20に示すように、上位計算機からハードディスク24と27にデータをライトすると同時に、ハードディスク25と28からデータをリードし、ハードディスク26からハードディスク27にデータをバックアップするといったことが可能になる。またハードディスク25に故障が発生した場合には例えば図21に示すように、ハードディスク24にはディスクアレイ制御部側の送信及び受信ポートが、またハードディスク26にはハードディスク27との送信及び受信ポートが残っている。

4

したがって、上位計算機からは故障したハードディスクを除いて、残りの全てのハードディスクに対してリード/ライトアクセスが行えるため、ディスクアレイ装置の連続運用が可能である。

【0005】

【発明が解決しようとする課題】従来の方法では、以下の課題がある。

【0006】すなわち、まず下位I/F制御部にSCSIを用いるディスクアレイ装置について述べる。ディスクアレイ制御部1が有する下位I/F制御部に1台以上の故障が発生した場合、その下位I/F制御部と接続しているハードディスクに対する上位計算機からのリード/ライトアクセスが完全に遮断されることが考えられる。また下位I/F制御部とハードディスクを相互に接続しているケーブルに、断線等の故障が発生した場合にもハードディスクに対するアクセスが停止すると考えられる。したがってこのような条件下では、故障の発生したI/Fチャネルを除いた構成でのディスクアレイ装置の連続運用は可能であるが、1台のハードディスクが使えないことによって、正常運用時に比較してディスクアレイ装置のリード/ライトのアクセス性能が低下すると予想される。

【0007】一方、下位I/F制御部にシリアルI/F制御部を用いるディスクアレイ装置について述べると、ディスクアレイ制御部及び各ハードディスクの持つ各々のシリアルI/F制御部がループ状に連なるため、ディスクアレイ制御部からハードディスクへのアクセス経路は、送信用及び受信用を1組にして2系統を保持している。これによって、いずれのハードディスクに故障が発生しても、残りの正常なハードディスクに対してアクセスを継続可能であるため、装置として高い信頼性がある。しかし、ディスクアレイ制御部側のシリアルI/F制御部に故障が発生した場合には、上位計算機とハードディスク間の通信経路を失うことから、ハードディスクのリード/ライトアクセスが事実上不可能となる。またディスクアレイ制御部側のシリアルI/F制御部と、ハードディスク側のシリアルI/F制御部間のケーブルに断線等の故障が一つ発生すると、2系統ある通信経路の内一つの送信用もしくは受信用の通信経路を失うことから、正常運用時に比較してディスクアレイ装置のリード/ライトのアクセス性能が低下すると予想される。

【0008】

【課題を解決するための手段】上記の目的を達成するため、ディスクアレイ制御部に予備の下位I/F制御部を搭載し、各下位I/F制御部が異なる2台ハードディスクと通信経路を持つようにデュアル化した通信手段（主にケーブルを意味する）、および各下位I/F制御部と通信経路の間に正常時に使用する主経路と故障時に使用する予備経路を選択可能な通信経路切替手段を設ける。

【0009】

(4)

特開平8-320768

5

【作用】上記の目的を達成する方法を用いると、下位 I/F 制御部及び通信経路に故障が発生していない場合、通信経路切替手段の主経路を有効、予備経路を無効に設定し、主経路と接続した通信経路の末端に接続してあるハードディスクに対してリード/ライトアクセスを行う。またある 1 台の下位 I/F 制御部が故障で使用不能になった場合、または運用中に通信経路が断線した場合には、故障した下位 I/F 制御部または断線した通信経路を持つ下位 I/F 制御部に対して右または左に隣り合う、全ての下位 I/F 制御部の通信経路切替手段の主経路を無効、予備経路を有効に設定し、予備経路と接続した通信経路の末端に接続してあるハードディスクに対してリード/ライトアクセスを継続して行う。ディスクアレイ装置の運用はこの状態で継続可能である。そして運用の終了後、一度ディスクアレイ装置の電源を落とし、故障した下位 I/F 制御部を新しい下位 I/F 制御部に交換する。この時、交換した下位 I/F 制御部に対して左に隣り合う、全ての下位 I/F 制御部の通信経路切替手段の主経路を再度有効、予備経路を無効に設定し、主経路と接続した通信経路の末端に接続してあるハードディスクに対してリード/ライトアクセスを行う。尚通信経路切替手段を別のドータボードに搭載し、ディスクアレイ制御部を構成する基盤上の下位 I/F 制御部とハードディスクの間にそのドータボードを付加することで、従来のディスクアレイ装置に対しても本方式の適用が図れる。

【0010】

【実施例】本発明の第一実施例を図 1 ないし図 8 を用いて説明する。

【0011】図 1 は本発明のディスクアレイ装置全体を表した図である。ディスクアレイ装置の中核となるディスクアレイ制御部 1 は、上位計算機との間でコマンド及びデータの授受を行う上位計算機 I/F 制御部 2、複数のハードディスク 13～17 との間でコマンド及びデータの授受を行う下位 I/F 制御部 7～12、各々の下位 I/F 制御部 7～12 と相互に接続し、下位 I/F 制御部とハードディスク間の通信経路（以下ケーブルと記す）46～55 の有効/無効の切替えを行うデュアル I/F スイッチコネクタ 40～45、コマンドのアドレス変換及びデータの集合分散制御を行うデータ転送制御部 5、上位計算機とハードディスク 13～17 間で転送するデータを一時的に格納保持するバッファメモリ 4、ハードディスク 13～17 にデータをライトする際にデータに付加するパリティを生成するパリティ生成回路 6、さらに以上の各構成要素を統括制御する MPU 3 から構成される。尚デュアル I/F スイッチコネクタ 40 はハードディスク 13 に、デュアル I/F スイッチコネクタ 45 はハードディスク 17 に対して 1 本の通信経路を備え、残りのデュアル I/F スイッチコネクタ 41～44 はハードディスク 13～17 に対して各々 2 本の通信経

6

路を備えている。また各ハードディスクは図 8 に示すようなケーブルを用いて、隣り合う 2 台のデュアル I/F スイッチコネクタの各々 1 つの通信ポートと接続している。さらにディスクアレイ装置運用開始時にディスクアレイ制御部 1 上の各ハードウェアをチェックする CUDG<Controller Unit Diagnostic>プログラムをフロッピディスク 19 で供給するため、ディスクアレイ制御部 1 にフロッピディスクドライブ（以下 FDD と記す）18 を接続している。

【0012】図 2～図 4 のフローチャートに基づき本発明のディスクアレイ装置の動作について説明する。

【0013】ディスクアレイ装置の電源を入れたと（図 2 のフローチャート 100）、FDD 18 に入っているフロッピディスク 19 からロードされた CUDG プログラムが起動し、ディスクアレイ制御部 1 の各構成要素である上位計算機 I/F 制御部 2、MPU 3、バッファメモリ 4、データ転送制御部 5、パリティ生成回路 6、各下位 I/F 制御部 7～12 及び各ハードディスク 13～17 のハードチェックを行う（同 101）。この時下位 I/F 制御部 7～12 のいずれかの障害の発生、または下位 I/F 制御部 7～12 とハードディスク 13～17 をつなぐケーブルの断線によって、ハードディスク 13～17 のハードチェックが正常に行われない場合は（同 102）、MPU 2 がエラー情報を上位計算機に報告する（同 103）。運用者はディスクアレイ装置の電源をいったん落とし、故障の認められた下位 I/F 制御部を交換するか、または断線していると思われるケーブルを交換する（同 105）。一方ディスクアレイ装置の電源投入時のハードチェックにおいて、いかなるハードエラーも認められなかった場合は（同 102）、図 5 の上側に示す下位通信経路の接続構成でディスクアレイ装置の運用を開始する。上位計算機は運用者の要求に応じてリード/ライトコマンドを送出し（同 106）、ディスクアレイ制御部 1 の MPU 2 がコマンドを解釈し、データ転送制御部 5 がリード時には各ハードディスク 13～17 に分散したデータの読み出し集合処理、ライト時には各ハードディスク 13～17 へデータの分散書き込み処理を実行する（同 107）。ディスクアレイ装置のリード/ライトアクセス実行中にいずれの下位 I/F 制御部、ケーブル及びハードディスクの故障がなく（同 108）、実行中のリード/ライトアクセスがまだ終了していない場合は（同 109）、引続きリード/ライトアクセスを実行する（同 107）。一方リード/ライトアクセスが終了している場合（同 109）、ディスクアレイ制御部 1 が上位計算機から次のリード/ライトアクセス要求を受け付ける。次のリード/ライトアクセス要求があるならば（同 110）上位計算機が次のリード/ライトコマンドをディスクアレイ制御部 1 に対して送出し（同 106）、ないならば（同 110）ディスクアレイ装置の運用を終了する（同 111）。

(5)

特開平8-320768

7

8

【0014】ここで例えば図1の下位I/F制御部9、ケーブル49またはハードディスク14のいずれかに故障が発生したと仮定する(同108)。リード時には(同112)データ転送制御部5が上位計算機にリードデータを転送する際に、読み出しデータからLRC<Longitudinal Redundancy Check>コードを生成し、読み出しデータが最後にハードディスクに書き込まれた際にデータ転送制御部5が読み出しデータ末尾に付加したLRCコードと比較する(図3のフローチャート115)。ディスクアレイ制御部1がLRCコードの不一致を確認したら(同116)、MPU3がデュアルI/Fスイッチコネクタ40、41を切替え(同117)、図5の上側に示す通信経路の接続構成から下側に示す通信経路の接続構成に変更し、同一データのリードを再実行する(同118)。再びデータ転送制御部5が上位計算機にリードデータを転送する際に、読み出しデータからLRCコードを生成し、読み出しデータが最後にハードディスクに書き込まれた際にデータ転送制御部5が読み出しデータ末尾に付加したLRCコードと比較する(同119)。再び上位計算機がLRCコードの不一致を確認したら(同120)、ハードディスク14または切替えた通信経路に故障が発生していると考えられるため(同121)、縮退モード(故障の認められたハードディスクを除いて、残りのハードディスクに対してリード/ライトアクセスを行う)にてディスクアレイ装置の運用を継続する(同122)。一方ディスクアレイ制御部1がLRCコードの一致を確認したら(同120)、下位I/F制御部8に故障またはケーブル49に断線が発生していると考えられるが(同123)、これらの故障箇所は無視されるため、ディスクアレイ装置のその後の運用には何も問題はない。したがってそのままディスクアレイ装置の運用を継続する(同124)。運用終了後ディスクアレイ装置の電源をいったん落とし(同125)、下位I/F制御部8またはケーブル49を含む図8で示したケーブルの交換を行う(同126)。一方ライト時でかつライト異常が発生した場合(同112)、MPUが管理情報として専用のメモリに持っているデータのマッピングテーブルに、ライトの異常実行に関する情報を書き込む(同113)。その後ライトしたデータをリードする時点になると(同114)、MPUがマッピングテーブルを参照し、異常の認められたI/FチャネルのデュアルI/Fスイッチコネクタ40、41を切替え、図5の上側に示す通信経路の接続構成から下側に示す通信経路の接続構成に変更し(図4のフローチャート127)、ライトしたと思われるデータのリードを実行する(同128)。読み出しデータからLRCコードを生成し、読み出しデータが最後にハードディスクに書き込まれた際にデータ転送制御部5が読み出しデータ末尾に付加したLRCコードと比較する(同129)。ディスクアレイ制御部1がLRCコードの不一致を確認し

10

20

30

40

50

たら(同130)、故障の発生箇所はハードディスク12と特定できるため(同131)、前述のリード時と同様に、縮退モードにてディスクアレイ装置の運用を継続する(同132)。一方、ディスクアレイ制御部1がLRCコードの一致を確認したら(同130)、下位I/F制御部8に故障またはケーブル49に断線が発生していると考えられるが(同133)、ハードディスク12からリードするデータはライトする前の旧データと考えられるため、残りのハードディスク13、15、16、17にライトしたデータ及びバリティから書き込みに失敗したデータを再生し、上位計算機に転送すると共にハードディスク12に再生データをライトする(同134)。尚前述の故障箇所は、ディスクアレイ装置のその後の運用には何も問題はないため、そのまま運用を継続する(同135)。運用終了後ディスクアレイ装置の電源を一度落とし(同136)、下位I/F制御部8またはケーブル49を含む図8で示したケーブルの交換を行う(同137)。

【0015】次に本発明で適用したデュアルI/Fスイッチコネクタ40～45の機能について説明する。

【0016】図6は図1のデュアルI/Fスイッチコネクタ40～45内部の詳細を表した図である。

【0017】デュアルI/Fスイッチコネクタ40～45は、図1の下位I/F制御部7～12とコネクタにより接続し、図1のディスクアレイ制御部1がハードディスク13～17に対して送出するコマンド及びデータの転送経路の入口、またハードディスク13～17がディスクアレイ制御部1に対して送出するコマンド及びデータの転送経路の出口となる下位I/F制御部コネクタ接続ポート60、下位I/F制御部及びケーブル断線等のハードエラーの発生によって、ディスクアレイ制御部1が送出する下位I/F制御部7～12とハードディスク13～17の通信経路を切替る制御信号を受ける、スイッチ切替(以下SWCTL<Switch Control>と記す)信号入力ポート61、下位I/F制御部7～12とハードディスク13～17の間の通信方向を切替る制御信号を受ける、通信方向切替(以下WREN<Write Enable>と記す)信号入力ポート62、ディスクアレイ装置の正常運用時に使用する主経路の有効/無効及び通信方向を切替る手段63、ディスクアレイ装置の下位I/F制御部7～12及びケーブル46～55の断線等のハードエラー時に使用する、予備経路の有効/無効及び通信方向を切替る手段64、SWCTL信号に従って、予備経路の有効/無効及び通信方向を切替る手段64のWREN信号の入力の有効/無効を切替る手段65、ハードディスク13～17とコネクタにより接続し、ハードディスク13～17がディスクアレイ制御部1に対して送出するコマンド及びデータの転送経路の入口、また下位I/F制御部7～12がハードディスク13～17に対して送出するコマンド及びデータの転送経路の出口となる

(6)

特開平8-320768

9

10

ハードディスクコネクタ接続ポート66、67で構成さ

*【0018】

れる。

*【表1】

表1 デュアルI/Fスイッチコネクタ制御信号仕様一覧

		SWCTL-N	WREN-N
MAIN	Write	偽	真
	Read		偽
SUB	Write	真	真
	Read		偽

【0019】表1に、デュアルI/Fスイッチコネクタ制御信号の一覧を示す。

【0020】図7のフローチャートに基づき本発明のディスクアレイ装置のデュアルI/Fスイッチコネクタの内部動作について説明する。

【0021】ディスクアレイ装置の電源を入ると（図7のフローチャート200）、図1のディスクアレイ制御部1のMPU3がデュアルI/Fスイッチコネクタ40～45のSWCTL信号を偽に設定する（同201）。したがって主経路の通信経路を用いてディスクアレイ装置の運用を開始する（同202）。まず上位計算機がリード/ライトコマンドをディスクアレイ制御部1に送出すると（同203）、リードコマンドを受けた場合（同204）、ディスクアレイ制御部1のMPU3がデュアルI/Fスイッチコネクタ40～45のWREN信号を偽に設定する（同205）。一方ライトコマンドを受けた場合は（同204）、WREN信号を真に設定する（同206）。以上の設定の後、ハードディスク13～17に対してリード/ライトアクセスを実行する（同207）。いずれの下位I/F制御部7～12、ケーブル46～55及びハードディスク13～17に故障が発生していなければ、前述の設定は変えずにリード/ライトアクセスを継続し、アクセス終了後ディスクアレイ制御部1が上位計算機から次のリード/ライトコマンドの要求を受けなければ（同211）、ディスクアレイ装置の運用は終了する（同212）。一方次のリード/ライトコマンドの要求を受ければ（同211）、引続き次のリード/ライトアクセスを実行する。しかし、リード/ライトアクセス実行中に下位I/F制御部7～12、ケーブル46～55及びハードディスク13～17のいずれかに故障が発生した場合には（同208）、ディスクアレイ制御部1のMPU3がデュアルI/Fスイッチコネクタ40～45のSWCTL信号を真に設定する（同209）。したがって予備経路を用いてディスクアレイ装置の運用を継続する。以後ディスクアレイ装置の運用を終了するまで、前述の設定は変更しない。

【0022】このように本方式のディスクアレイ装置は、従来の下位I/F制御部とハードディスクを一対一対応に接続している従来のディスクアレイ装置に比較して、予備の下位I/F制御部を付加することで、1台の下位I/F制御部の故障発生もしくは1本のケーブルの断線がディスクアレイ装置の運用に全く影響を及ぼさ

いという特徴を持っている。これはつまり本方式を適用することで、ディスクアレイ装置の下位I/F制御部、ハードディスク及びそれらを相互に接続するケーブルを含む下位I/Fチャネルに対する耐故障の信頼性を高めることができることを意味している。また図1の下位I/F制御部7～12にソケット式のものをを用いることで、ディスクアレイ制御部1を構成する基盤を交換せずに故障の発生した下位I/F制御部のみを交換することで、修理コストを低減できる。さらにデュアルI/Fスイッチコネクタ40～45とハードディスク13～17の間に使用するケーブルの形状も、2本ケーブルの一端を一つに束ねていることを除けば従来のケーブルと同等であるため、取り扱いには用意である。

【0023】第一実施例ではディスクアレイ制御部に対して新たな予備の下位I/F制御部を付加する方法を提示したが、特にハードウェアのコストを抑えてディスクアレイ装置を構成する必要がある場合には、ディスクアレイ制御部にすでに搭載している本来バックアップ機器を接続するために使用する下位I/F制御部を予備の下位I/F制御部として流用する方法が考えられる。この方法について次の第二実施例で述べる。

【0024】本発明の第二実施例を図9、図10を用いて説明する。

【0025】図9は従来のディスクアレイ装置が外部機器の接続用に持っている下位I/F制御部を、第一実施例で述べた本発明のディスクアレイ装置に付加した予備の下位I/F制御部に流用したディスクアレイ装置全体を表した図である。ディスクアレイ制御部1のハードウェア構成は第一実施例で述べたディスクアレイ装置と同様であるが、図6で示したデュアルI/Fスイッチコネクタ40、45の各々の二つのハードディスクコネクタ接続ポートの内1つが空いているため、ここにバックアップ機器用I/Fコネクタ56、57を設ける。ここではバックアップ機器用I/Fコネクタ56にバックアップ機器を接続して、ハードディスク13～17のデータをバックアップする方法について説明する。まずディスクアレイ装置がリード/ライトアクセス実行中に（図10のフローチャート300）上位計算機がハードディスク13～17のデータのバックアップを要求すると（同301）、データ転送制御部5がハードディスク13～17からデータを収集し一時データバッファ4へ転送保持する（同302）。下位I/F制御部、使用中のケー

(7)

特開平8-320768

11

ブルに故障が発生していない間は、主経路であるケーブル47、49、51、53、55がディスクアレイ制御部1とハードディスク13～17間の通信経路になるが、バックアップ機器はバックアップ機器用I/Fコネクタ56を介してデュアルI/Fスイッチコネクタ40の主経路を利用できるため、下位I/F制御部7をバックアップ機器専用の下位I/F制御部として使うことができる。したがって、データバッファ4に保持しているバックアップデータを即バックアップ機器に転送する(同309)。バックアップ完了後(同310)、再びハードディスク13～17に対するリード/ライトアクセスを継続する。一方いずれの下位I/F制御部、使用中のケーブルに故障が発生した場合には(同303)、デュアルI/Fスイッチコネクタの予備経路とケーブル46がハードディスク13との通信経路になるため、ハードディスク13に対するリード/ライトアクセスを優先する。上位計算機がハードディスク13～17のデータのバックアップを要求すると(同301)、データ転送制御部5がハードディスク13～17からデータを収集し一時データバッファ4へ転送保持する(同302)。ハードディスク13に対する次のリード/ライトアクセス要求が来ない間に(同304)、デュアルI/Fスイッチコネクタ40を主経路に切替え(同305)、データバッファ4に保持しているバックアップデータをバックアップ機器に転送する(同306)。バックアップ完了後(同307)、デュアルI/Fスイッチコネクタ40を予備経路に切替え(同308)、再びハードディスク13～17に対してリード/ライトアクセスを継続する。

【0026】第二実施例で示したバックアップ方法を利用すると、正常時のみならずリード/ライトアクセス中に下位I/F制御部8～12、ケーブル47、49、51、53、55に故障が発生した場合にもハードディスク13～17のデータのバックアップができるため、データ喪失の発生を抑えることができる。しかし、バックアップに使用中の下位I/F制御部7、バックアップ機器用I/Fコネクタ56が故障する可能性も考えられるため、この問題に対処するためにハードディスク13～17から転送したバックアップデータをバッファメモリ4に保持する時にバックアップデータのコピーをつくり、バックアップ処理が完了するまでそのコピーデータをバッファメモリ4に保持しておく。したがって、バックアップデータをバックアップ機器19に転送中に前述の故障が発生してもコピーデータはバッファメモリ4に残っているため、バックアップ機器19をバックアップ機器用I/Fコネクタ57につなぎ替えることでバックアップを再実行することができる。この場合もハードディスク13～17に対するリード/ライトアクセスを優先するため、ハードディスク17に対する次のリード/

12

チコネクタ45を主経路から予備経路に切替えバックアップを実行し、バックアップ完了後デュアルI/Fスイッチコネクタ45を予備経路から主経路に戻し、再びハードディスク13～17に対してリード/ライトアクセスを継続する。ここで挙げた方式は、バッファメモリ4にバックアップデータのコピーを保持できる空き容量のある場合には特に勧められるが、空き容量に余裕がない場合も考えられる。この問題を考慮したバックアップ方法を次の第三実施例で述べる。

【0027】本発明の第三実施例を図11を用いて説明する。

【0028】図11は第二実施例で述べたディスクアレイ装置を変形させたディスクアレイ装置全体を表した図である。ディスクアレイ制御部1のハードウェア構成は第二実施例で述べたディスクアレイ装置と同様であるが、デュアルI/Fスイッチコネクタ40、45の各々1つの空いているハードディスクコネクタ接続ポートにそれぞれ1本のケーブルを接続し、それら2本のケーブルの終端にMPUで操作するバックアップデータ転送経路切替スイッチ58を設ける。そしてバックアップデータ転送経路切替スイッチ58から出ているバックアップ機器用I/Fコネクタ59にバックアップ機器19を接続する。ここで述べるディスクアレイ装置のバックアップ動作は第二実施例で示したディスクアレイ装置とほぼ同様であるが、バックアップデータのコピーを作りバッファメモリ4に保持しておく必要はない。それはつまり、下位I/F制御部7でバックアップを実行中に、下位I/F制御部7及びそれとバックアップデータ転送経路切替スイッチ58をつないでいるケーブルに故障が発生しても、MPUがバックアップデータ転送経路切替スイッチ58を図11に示すポジションCからAに切替えることで、下位I/F制御部12を用いてバックアップデータ転送を継続することができるからである。また上位計算機からバックアップの要求を受けていない間は、MPUがバックアップデータ転送経路切替スイッチ58を図11に示すポジションBに設定するので、バックアップ機器19の取り外しや交換はディスクアレイ装置の電源を落とさずに自由に行うことができる。本ディスクアレイ装置は、バックアップデータ転送経路切替スイッチ58と僅かなケーブルを付加することで、いずれの下位I/Fチャネルの故障発生に関わらず、正常運用時と同様にハードディスクのデータをバックアップできる特徴を持っている。

【0029】特に第二、第三実施例で述べたディスクアレイ装置は、予備の下位I/F制御部に本来外部機器専用の下位I/F制御部を流用することから、従来のディスクアレイ装置とほとんど変わらないハードウェアコストで、下位I/Fチャネルの耐故障性を高めことができる方法として勧められる。

【0030】ところで、従来のディスクアレイ装置のハ

(8)

特開平8-320768

13

ードウェア構成を変えずに、拡張ボード（以下ドーたボードと記す）を付加することで、第一から第三実施例で提案したのと同様な下位I/Fチャネルの耐故障性の高いディスクアレイ装置を実現する方法も考えられる。この方法について次の第四実施例で述べる。

【0031】本発明の第四実施例を図12を用いて説明する。

【0032】図12はデュアルI/Fスイッチコネクタ及びバックアップデータ転送経路切替えスイッチを搭載したドーたボードを付加したディスクアレイ装置全体図である。ディスクアレイ制御部1を構成する要素は従来のディスクアレイ装置と全く同様であり、第一から第三実施例までディスクアレイ制御部1を構成する基盤の内部に搭載していたデュアルI/Fスイッチコネクタ40～45及びバックアップデータ転送経路切替えスイッチ58を一枚のドーたボード68に搭載し、ディスクアレイ制御部1とハードディスク13～17の間に設置し、図8で示した形状のケーブルでデュアルI/Fスイッチコネクタ40～45とハードディスク13～17を接続することで、従来のディスクアレイ装置を下位I/Fチャネルの耐故障性の高いディスクアレイ装置へ容易にアップグレードできることを示している。

【0033】第一から第四実施例は下位I/F制御部にSCSIを用いたディスクアレイ装置を元にしていて、下位I/Fチャネルの耐故障性の高いディスクアレイ装置を提供する本発明は、次世代のI/F制御部として注目されているシリアルI/Fを用いたディスクアレイ装置に対しても適用できるものである。本発明を適用したシリアルI/Fを用いたディスクアレイ装置について次の第五実施例で述べる。

【0034】第五実施例を図13を用いて説明する。

【0035】図13はシリアルI/Fを用いた本発明のディスクアレイ装置全体図である。ディスクアレイ制御部69には、図17に示したシリアルI/Fを用いた従来のディスクアレイ装置のディスクアレイ制御部29に、シリアルI/F制御部22、シリアルI/F制御部22を制御する専用のシリアルI/F制御用MPU23と同等の機能を持つ予備用I/F制御部として、シリアルI/F制御部70、シリアルI/F制御部70を制御する専用のシリアルI/F制御用MPU71を持つ。またデータ転送制御部5とシリアルI/F制御部22またはシリアルI/F制御部70に接続を切替える通信経路切替えスイッチ72、さらにシリアルI/F制御部22とハードディスク24～28間及びシリアルI/F制御部とハードディスク24～28間の通信経路の有効/無効を切替えるデュアルI/Fスイッチコネクタ73をディスクアレイ制御部69内部に持つ。本発明のディスクアレイ装置は予備用I/F制御部を備えることで、リード/ライトアクセス実行中にシリアルI/F制御部22に故障が発生した場合にも、リード/ライトアクセスの

14

異常実行によって発生するデータの不一致を基に、MPU3が通信経路切替えスイッチ72及びデュアルI/Fスイッチコネクタ73を予備のシリアルI/F制御部70が使用できるように切替え、実行中のリード/ライトアクセスを継続することができる。尚ディスクアレイ装置の運用が終了したら、一度ディスクアレイ装置の電源を落とし、ディスクアレイ制御部69を構成するメインボードを交換するか、もしくはソケット式のシリアルI/F制御部を用いていれば、シリアルI/F制御部のみを交換する。

【0036】次にシリアルI/Fを用いた本発明のディスクアレイ装置に適用したデュアルI/Fスイッチコネクタ73の機能について説明する。

【0037】図14は図13のデュアルI/Fスイッチコネクタ73内部の詳細を表した図である。デュアルI/Fスイッチコネクタ73は、図13のシリアルI/F制御部22、70とコネクタにより接続し、図13のディスクアレイ制御部69がハードディスク24～28に対して送出するコマンド及びデータの転送経路の入口、またハードディスク24～28がディスクアレイ制御部69に対して送出するコマンド及びデータの転送経路の出口となるディスクアレイ制御部側シリアルI/F制御部通信入出力ポート75～82、シリアルI/F制御部22のハードエラーの発生時に、シリアルI/F制御部22とハードディスク24～28から、シリアルI/F制御部70とハードディスク24～28の通信経路に切替えるため、ディスクアレイ制御部69が送出する制御信号を受けるスイッチ切替（以下SWCTLと記す）信号入出力ポート74、ディスクアレイ装置の正常運用時に使用する主経路の有効/無効及び通信方向を切替える手段83～86、ディスクアレイ装置のシリアルI/F制御部22のハードエラー時に使用する、予備経路の有効/無効及び通信方向を切替える手段87～90、ハードディスク24～28とコネクタにより接続し、ハードディスク24～28がディスクアレイ制御部69に対して送出するコマンド及びデータの転送経路の入口、またシリアルI/F制御部22またはシリアルI/F制御部70がハードディスク24～28に対して送出するコマンド及びデータの転送経路の出口となるハードディスクコネクタ側シリアルI/F制御部通信入出力ポート91～94で構成される。ディスクアレイ装置の電源投入後（図15のフローチャート400）、ディスクアレイ制御部69のMPU3がSWCTL信号を偽に設定する（同401）。したがって主経路の有効/無効及び通信方向を切替える手段83～86は有効になり、予備経路の有効/無効及び通信方向を切替える手段87～90は無効になる（同402）。通信経路の確定後上位計算機がディスクアレイ制御部69にリード/ライトコマンドを送出し（同403）、ハードディスク24～28に対するリード/ライトアクセスを開始する（同404）。その後リ

(9)

特開平8-320768

15

ード／ライトアクセス中にシリアルI／F制御部22に故障が発生しなければ(同405)、現在実行中のリード／ライトアクセス終了後(同406)、上位計算機が次のコマンドを送出している場合、引続きリード／ライトアクセスを実行し(同407)、コマンド要求がなければディスクアレイ装置の運用を終了する(同408)。したがって正常運用時は主経路のみを用いることになる。一方リード／ライトアクセス中にシリアルI／F制御部22に故障が発生すると(同405)、ディスクアレイ制御部69のMPU3がSWCTL信号を真に設定する(同409)。したがって主経路の有効／無効及び通信方向を切替る手段83～86は無効になり、予備経路の有効／無効及び通信方向を切替る手段87～90は有効になる(同410)。尚その後上位計算機が次のコマンドを送出している場合、引続きリード／ライトアクセスを実行するが(同407)、ディスクアレイ装置の運用を終了するまで前述の設定は変更しない。

【0038】第五実施例で述べた本発明のディスクアレイ装置は、ディスクアレイ制御部の下位I／F制御部にシリアルI／Fを用いた従来のディスクアレイ装置の信頼性の問題を補う方式として提案できるものである。つまり各々のハードディスクの故障には2組ある送信受信ポートでディスクアレイ制御部との通信経路を常に確保できるが、コマンド及びデータをハードディスクに対して送信受信するディスクアレイ制御部本体の送信受信ポートでは故障を回避することが出来ない。

【0039】

【発明の効果】本発明によれば、下位I／F制御部、ケーブルの故障によるディスクアレイ装置の縮退モードによる運用(故障したI／Fチャネルのハードディスクを除いた、残りのハードディスクによるディスクアレイ装置の運用)が発生しないため、特にリードアクセス中にデータ転送性能の劣化のない連続運用が提供できる。

【0040】従来のディスクアレイ装置に下位I／Fチャネルの故障を回避する回路を搭載したドータボードを付加することで、装置の信頼性を高めるアップグレード方法を提供できる。

【0041】従来の下位I／F制御部にシリアルI／F制御部を用いたディスクアレイ装置に、下位I／Fチャネルの耐故障性を高める方法を提供できる。

【図面の簡単な説明】

【図1】本発明のディスクアレイ装置のブロック図。

【図2】本発明のディスクアレイ装置の動作を表したフローチャート。

【図3】本発明のディスクアレイ装置の動作を表したフローチャート。

【図4】本発明のディスクアレイ装置の動作を表したフローチャート。

【図5】実施例1における本発明のディスクアレイ装置の下位通信経路接続構成の説明図。

16

【図6】本発明のディスクアレイ装置に適用したデュアルI／Fスイッチコネクタの詳細を表した説明図。

【図7】デュアルI／Fスイッチコネクタの内部動作を表したフローチャート。

【図8】本発明のディスクアレイ装置に適用したハードディスクとデュアルI／Fスイッチコネクタを接続するケーブルの外観を表した説明図。

【図9】本発明のディスクアレイ装置にバックアップ装置を付加した状態を表した説明図。

10 【図10】バックアップ装置を付加した本発明のディスクアレイ装置のバックアップ動作を表したフローチャート。

【図11】本発明のディスクアレイ装置にバックアップ装置を付加した状態を表した説明図。

【図12】従来のディスクアレイ装置にデュアルI／Fスイッチコネクタを搭載したドータボードを適用した状態を表した説明図。

【図13】本発明のディスクアレイ装置(シリアルI／F使用)の説明図。

20 【図14】本発明のディスクアレイ装置(シリアルI／F使用)に適用したデュアルI／Fスイッチコネクタの詳細を表した説明図。

【図15】デュアルI／Fスイッチコネクタ(シリアルI／F使用)の内部動作を表したフローチャート。

【図16】従来のディスクアレイ装置の説明図。

【図17】パリティ生成及び喪失データ再生の仕組の説明図。

【図18】従来のディスクアレイ装置(シリアルI／F使用)の説明図。

30 【図19】シリアルI／F(SSAI／F)の詳細な構成を表した説明図。

【図20】従来のディスクアレイ装置(シリアルI／F使用)のハードディスクの故障時のアクセス状態を表した説明図。

【図21】従来のディスクアレイ装置(シリアルI／F使用)の複数データ転送実行状態を表した説明図。

【符号の説明】

- 1…ディスクアレイ制御部、
- 2…上位計算機I／F制御部、
- 40 3…MPU、
- 4…バッファメモリ、
- 5…データ転送制御部、
- 6…パリティ生成回路、
- 7～12…下位I／F制御部、
- 13～17…ハードディスク、
- 18…フロッピディスクI／F制御部、
- 19…フロッピディスクドライバ、
- 40～45…デュアルI／Fスイッチコネクタ、
- 46～55…ケーブル。

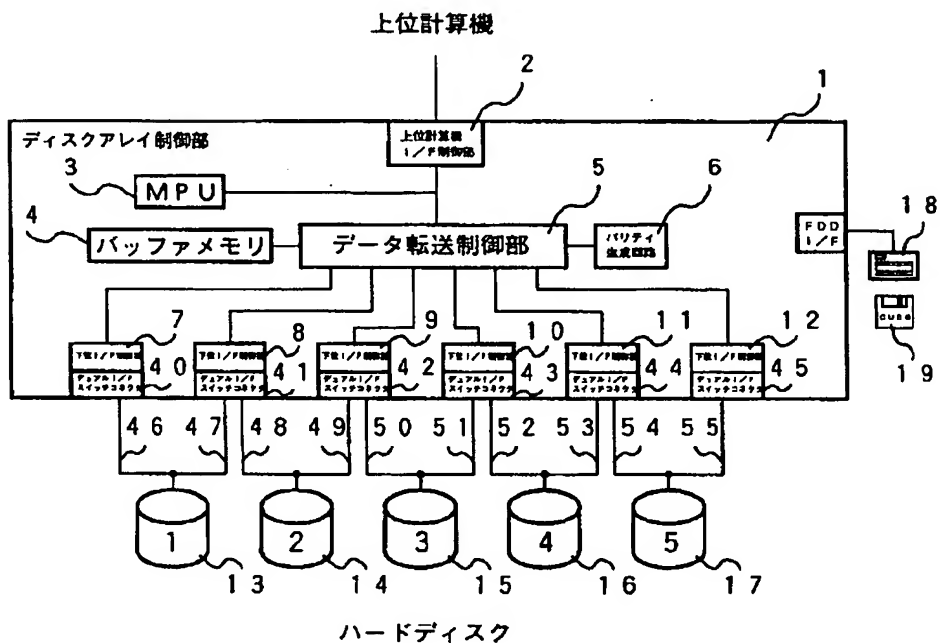
50

(10)

特開平8-320768

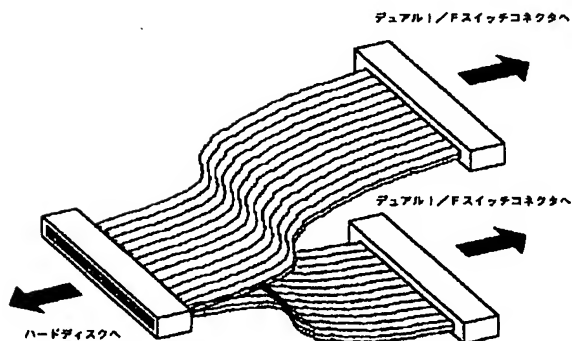
【図1】

図 1



【図8】

図 8



【図17】

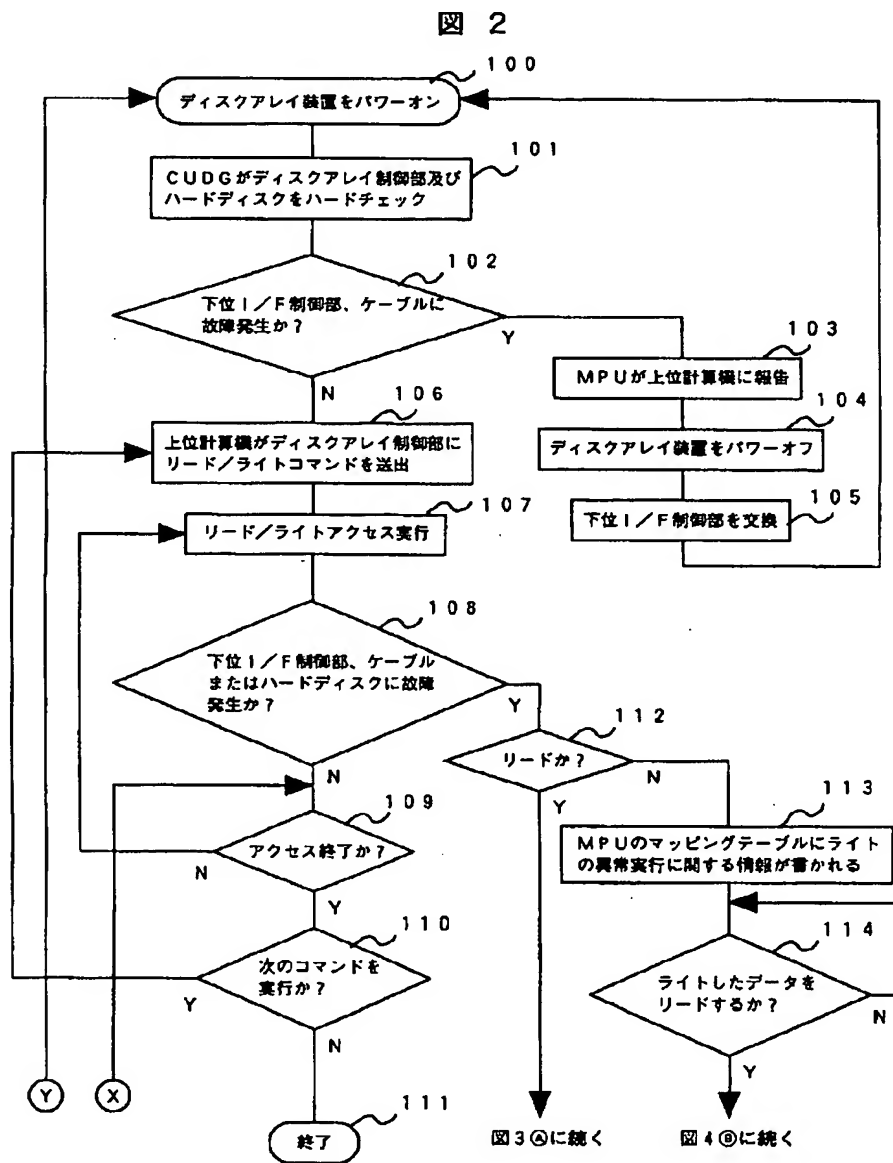
図 17

パリティ生成: $P = D1 \oplus D2 \oplus D3 \oplus D4 \oplus D5$ D3を喪失: ~~D3~~D3を再生: $D3 = D1 \oplus D2 \oplus D4 \oplus D5 \oplus P$

(11)

特開平8-320768

【図2】

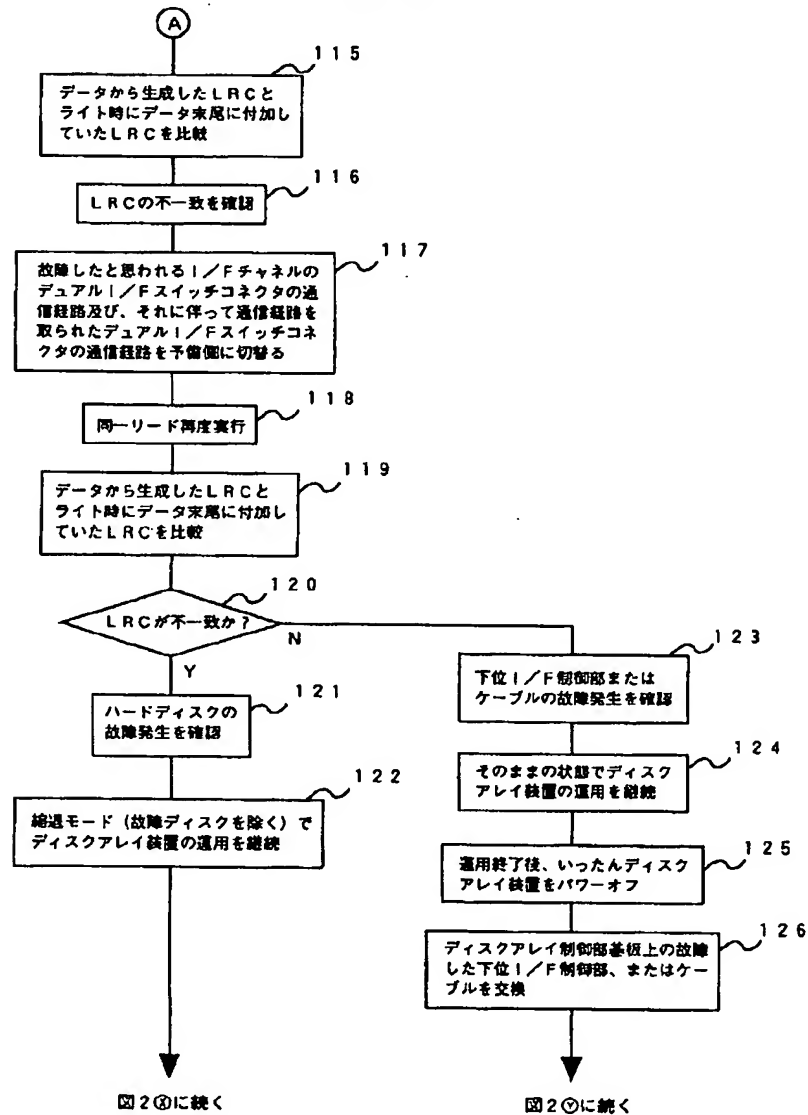


(12)

特開平 8-320768

【図 3】

図 3

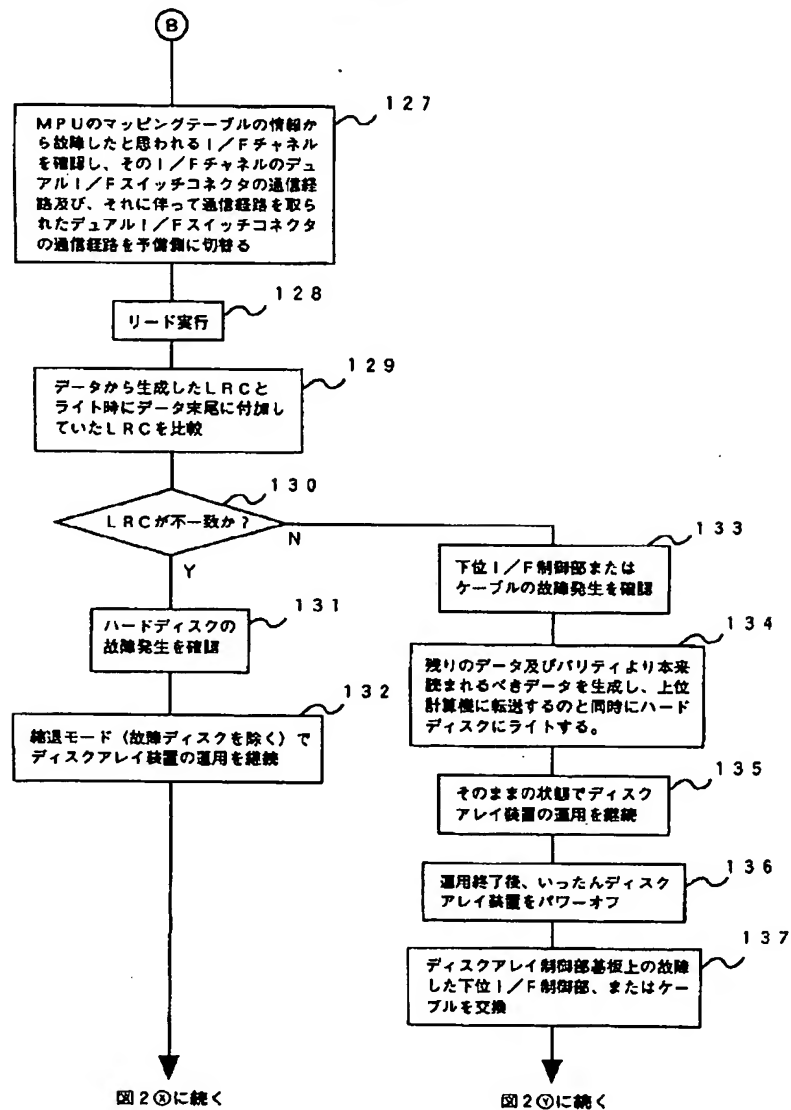


(13)

特開平 8-320768

【図 4】

図 4



(14)

特開平8-320768

【図5】

【図14】

図 5

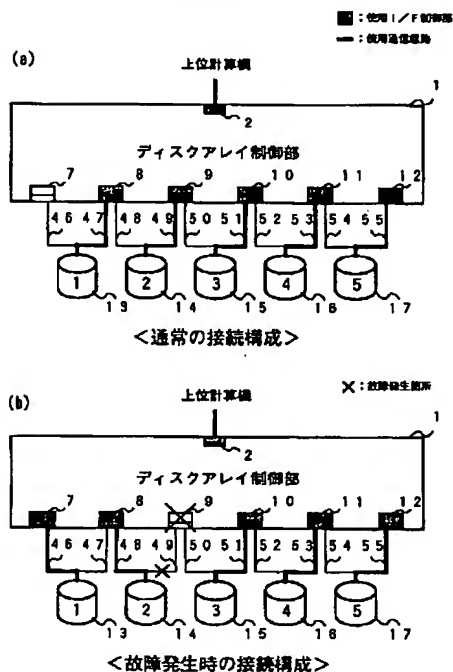
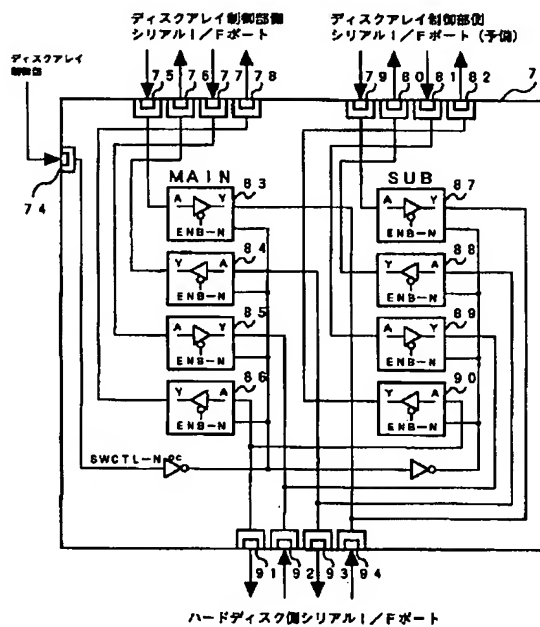
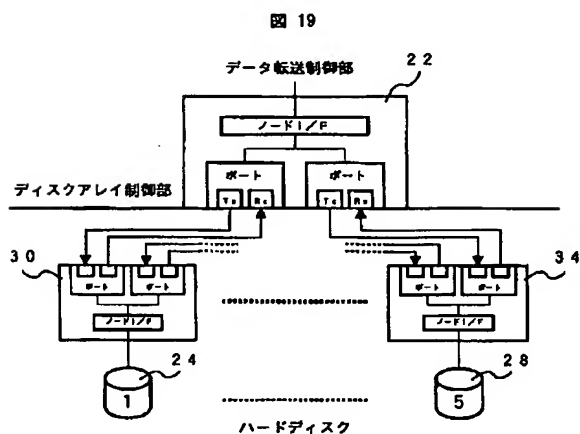


図 14



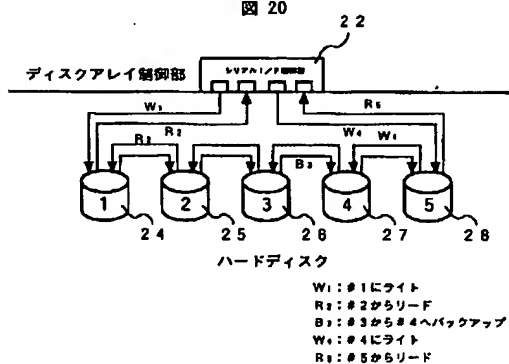
【図20】

【図19】



Tx : 送信ポート
Rx : 受信ポート

図 20

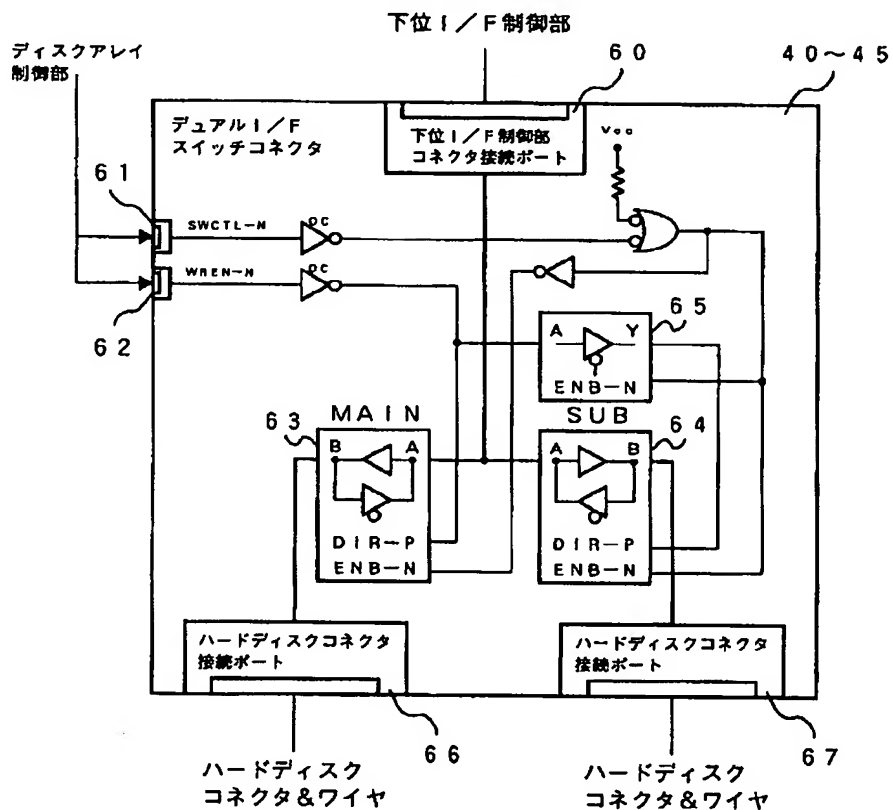


(15)

特開平 8-320768

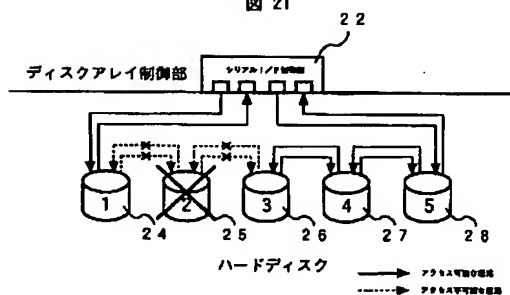
【図 6】

図 6



【図 21】

図 21

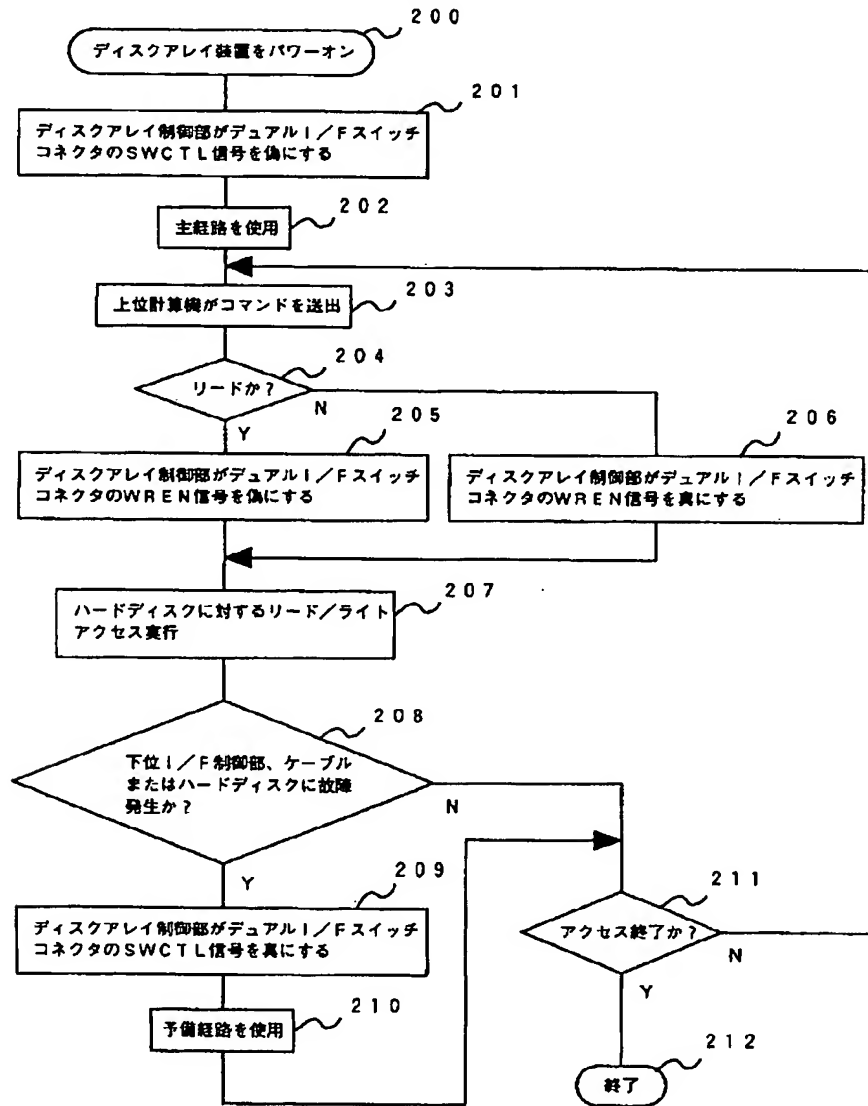


(16)

特開平8-320768

【図7】

図 7

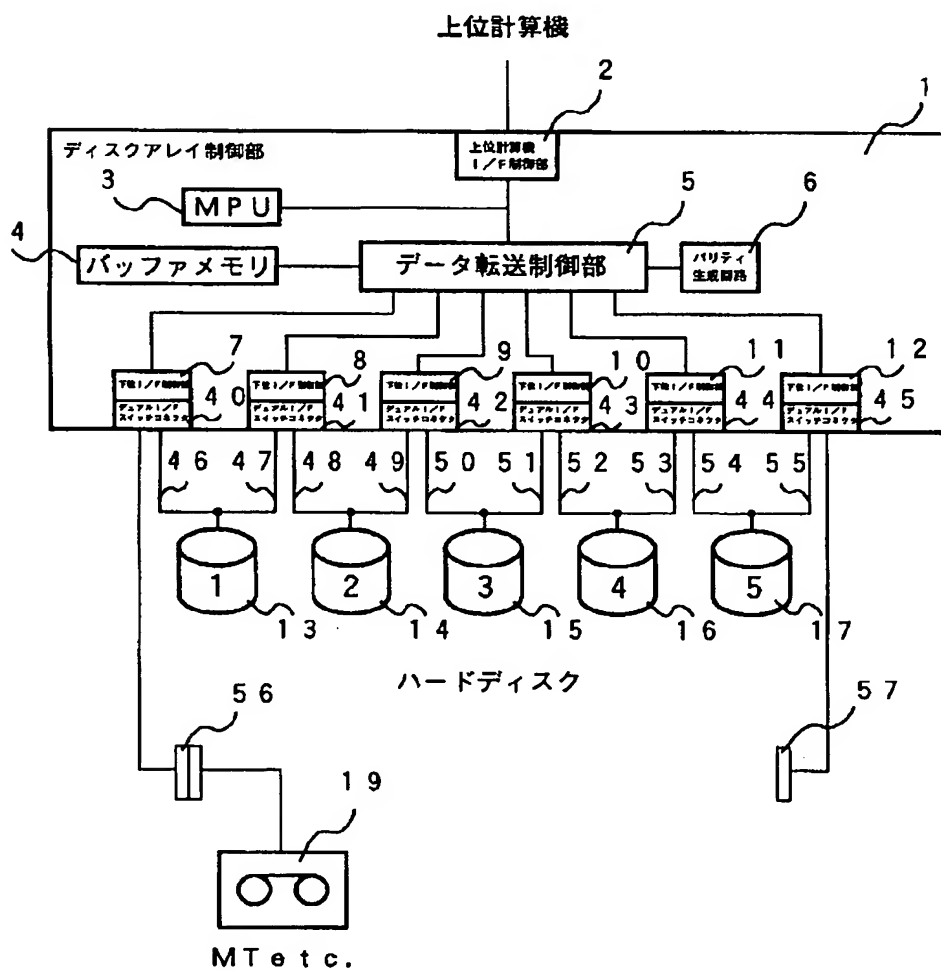


(17)

特開平8-320768

【図9】

図 9

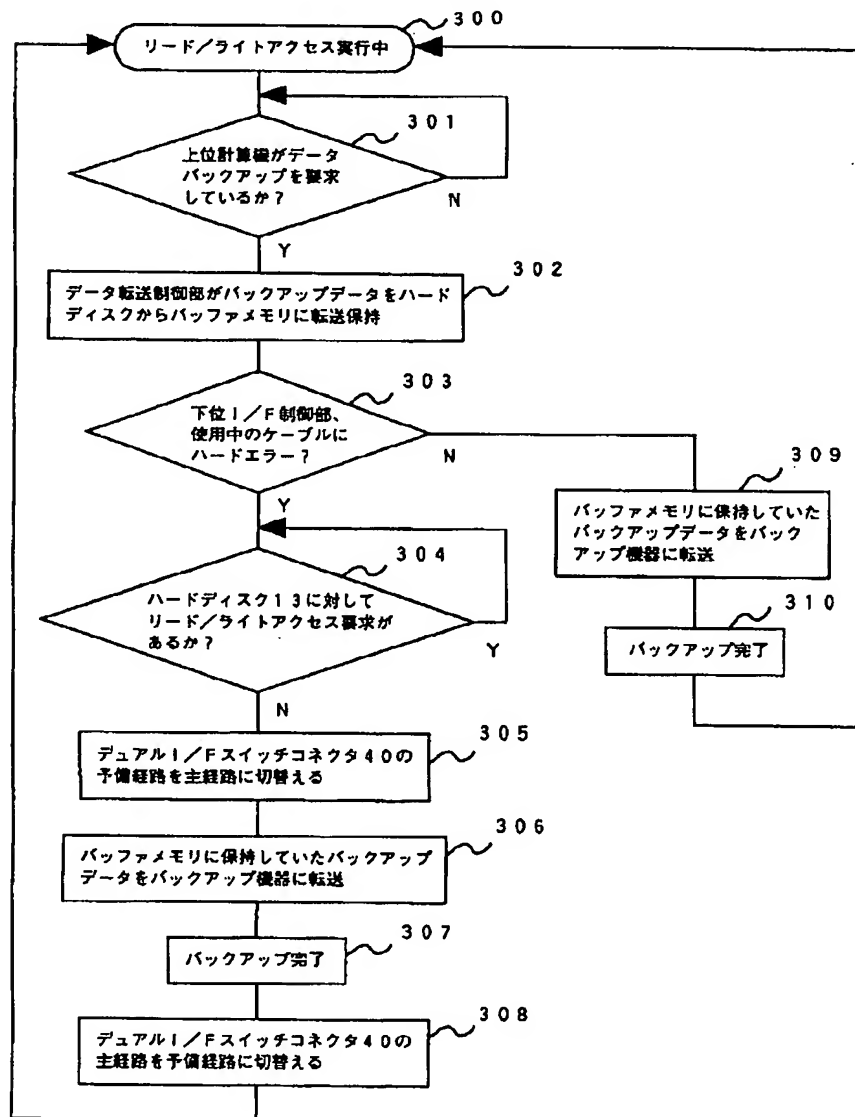


(18)

特開平8-320768

【図10】

図 10

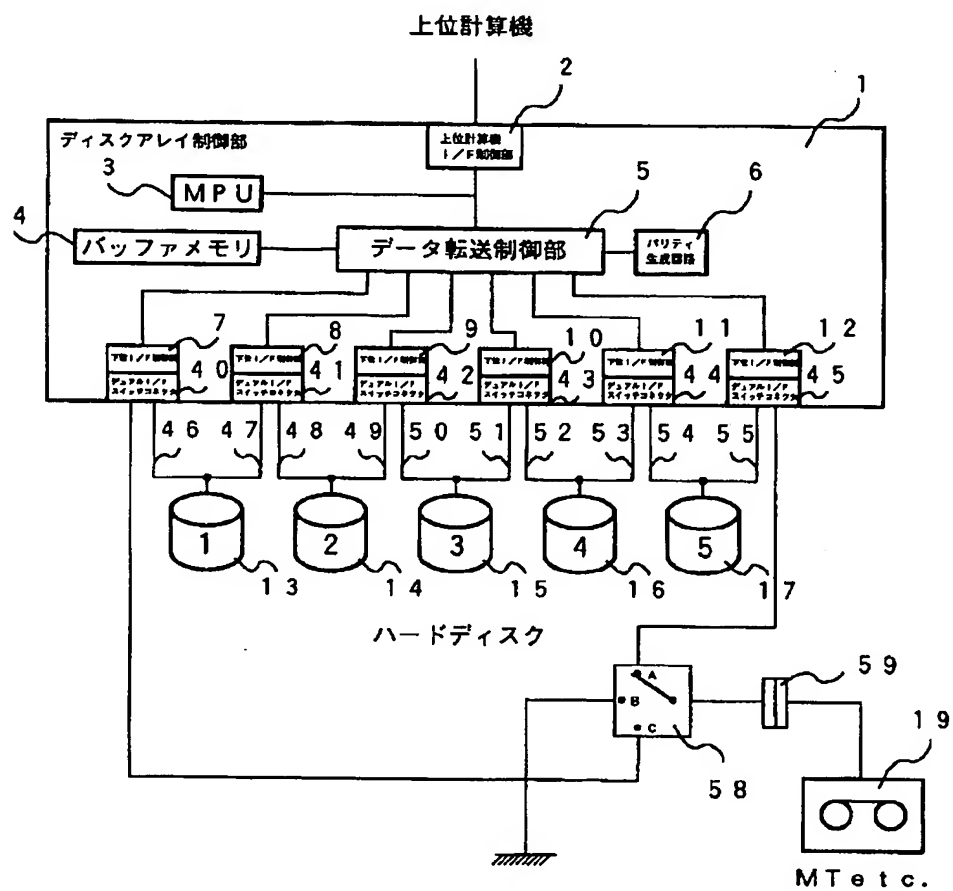


(19)

特開平8-320768

【図11】

図 11

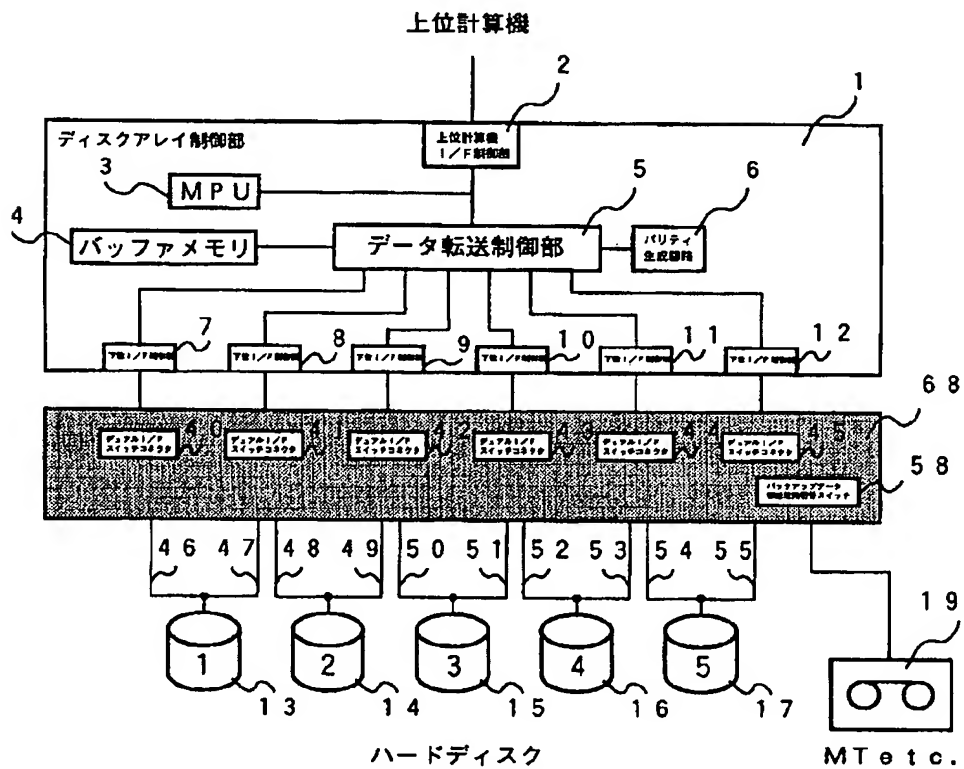


(20)

特開平 8-320768

【図 12】

図 12

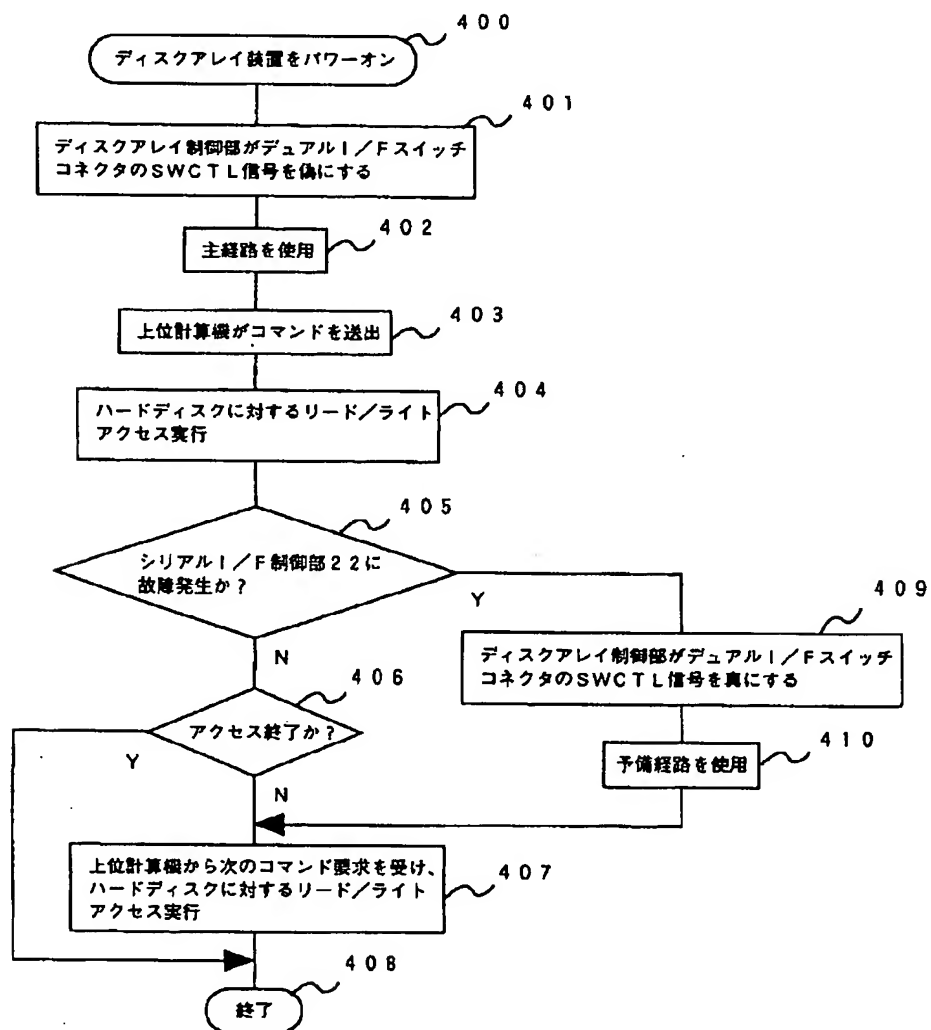


(22)

特開平 8-320768

【図15】

図 15



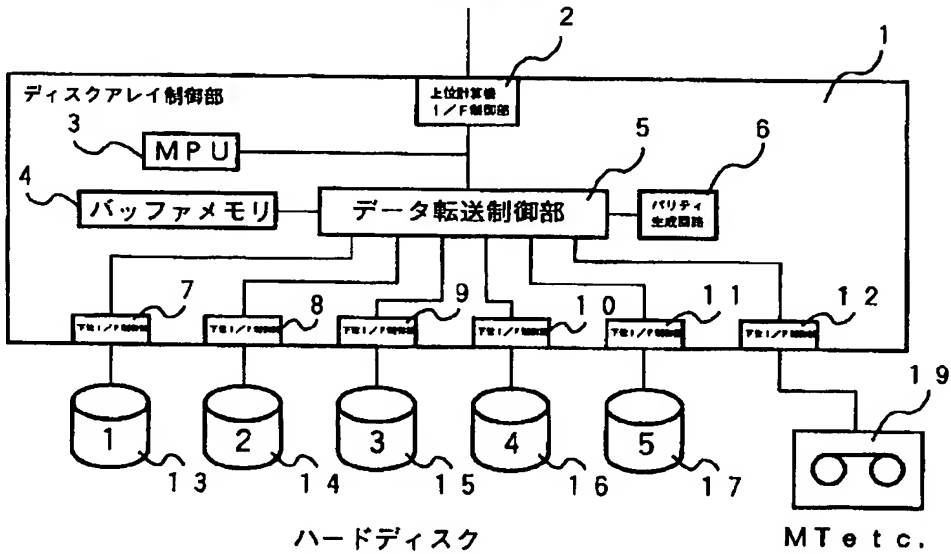
(23)

特開平 8 - 3 2 0 7 6 8

【図 16】

図 16

上位計算機

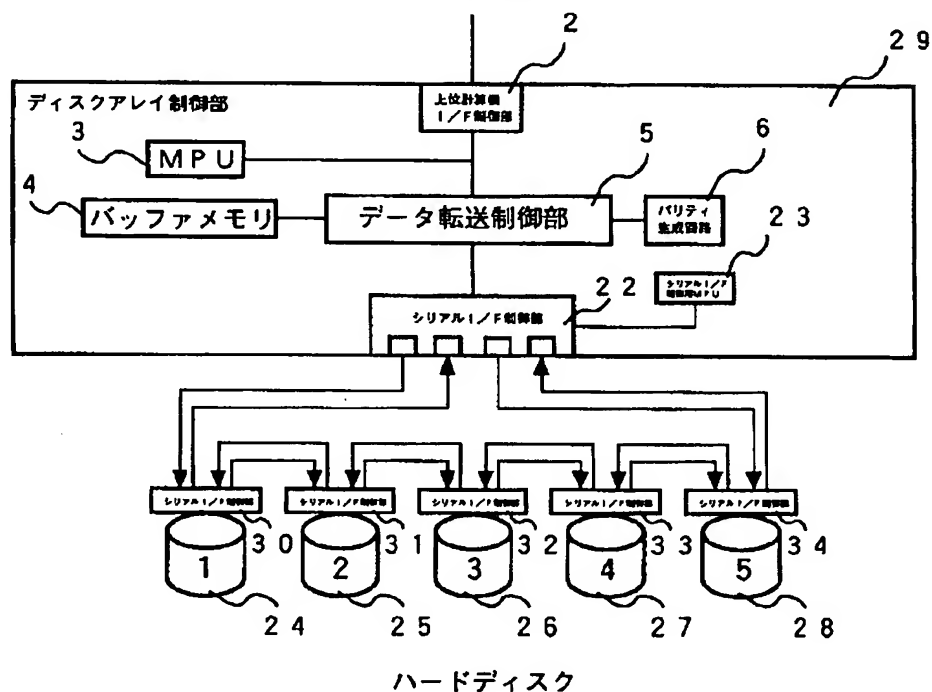


(24)

特開平 8-320768

【図18】

図 18
上位計算機



フロントページの続き

(72)発明者 本田 聖志

神奈川県川崎市麻生区王禅寺1099番地株式
会社日立製作所システム開発研究所内

(72)発明者 岩崎 秀彦

神奈川県小田原市国府津2880番地株式会社
日立製作所ストレージシステム事業部内